Enfoque UTE, V.7-Sup.1, Feb.2017, pp.172 - 185
http://ingenieria.ute.edu.ec/enfoqueute/
e-ISSN: 1390-6542 / p-ISSN: 1390-9363
Recibido (Received): 2017/01/09
Aceptado (Accepted): 2017/02/24

Estimating Sample Size for Usability Testing

(Estimación del tamaño de la muestra para pruebas de usabilidad)

Alex Cazañas, Andre de San Miguel [1], Esther Parra [2]

Abstract:
One strategy used to assure that an interface meets user requirements is to conduct usability testing. When conducting such testing one of the unknowns is sample size. Since extensive testing is costly, minimizing the number of participants can contribute greatly to successful resource management of a project. Even though a significant number of models have been proposed to estimate sample size in usability testing, there is still not consensus on the optimal size. Several studies claim that 3 to 5 users suffice to uncover 80% of problems in a software interface. However, many other studies challenge this assertion. This study analyzed data collected from the user testing of a web application to verify the rule of thumb, commonly known as the "magic number 5". The outcomes of the analysis showed that the 5-user rule significantly underestimates the required sample size to achieve reasonable levels of problem detection.

Keywords: usability testing; problem discovery; sample size

Resumen:
Una de las estrategias empleadas para asegurar que una interfaz satisfaga los requerimientos del usuario es la aplicación de pruebas de usabilidad. Cuando se aplican tales pruebas una de las incógnitas a despejar es el tamaño de la muestra. Dado que, realizar pruebas en forma extensiva es costoso, reducir al mínimo el número de participantes puede contribuir en gran medida a una gestión exitosa de los recursos de un proyecto. Aunque se han propuesto un número significativo de modelos para estimar el tamaño de la muestra en una prueba de usabilidad, todavía no hay consenso acerca del tamaño óptimo. Varios estudios afirman que de 3 a 5 usuarios son suficientes para descubrir el 80% de los problemas en la interfaz de un sistema. Sin embargo, otros tantos estudios cuestionan la validez de esta aseveración. El presente estudio analizó los datos obtenidos de las pruebas de usuario de una aplicación web para verificar la regla práctica conocida comúnmente como "el número mágico 5". El resultado del análisis mostró que la regla de los 5 usuarios subestima significativamente el tamaño de la muestra requerida para conseguir niveles razonables de detección de problemas.

Palabras clave: pruebas; usabilidad; muestra; descubrimiento; problemas

1. Introduction

The degree by which a software product meets user expectations in terms of its ease of use, effectiveness and efficiency, reflects the accomplishment of the intended usability goals of a software development project. Conducting usability testing is central to the realization of such goals because

[1] University of Queensland, Brisbane – Australia ({alex.cazanasgordon, andre.desanmiguel } @uqconnect.edu.au)
[2] Escuela Politécnica Nacional, Quito – Ecuador (esther.parra@epn.edu.ec)

it provides an effective tool to ensure that defects affecting usability are detected before the release of a new product.

In all projects, efficient resource allocation is essential, particularly when the cost of resources is high. Since increasing the amount of user testing sessions directly impacts the project cost, budget constraints will limit the ability to conduct exhaustive user testing. Consequently, from an economical point of view, it is important to ensure that the benefit gained by additional testing is greater than the incurred costs.

Determining the minimum number of participants that exposes most problems in a usability test is a problem that has generated a considerable amount of research and debate during the past two decades. In the early 1990's, Virzi (1992), Nielsen and Landauer (1993), and Lewis (1994) were the first to publish methods to estimate the size of the smallest sample required to achieve a target proportion of problems discovered in a usability test. Their research, based on empirical data and statistical modelling, allowed them to make three outstanding claims:

1) Most problems are discovered by the first three to five participants.
2) The increment in problem discovery after five participants is minimal.
3) ROI of usability testing can be maximized by minimizing the sample size.

Since their publication, these claims also known as "4±1" or "magic number 5" have generated a great deal of discussion in the usability community, so much so that at the Computer-Human Interaction conference in 2003, a panel was dedicated to discuss this matter (Bevan, et al., 2003).

This paper will recap on, and review the 4±1 model for estimating the sample size required to obtain a proportion of problem discovery of at least 80% in the testing of a web interface. Furthermore, the outcomes of a permutation test will be presented to investigate the effect of small samples on the estimation of the discovery rate.

1.1. Background

Three studies were considered for the provision of a suitable base calculation template to estimate sample size: (Virzi, 1992), (Nielsen & Landauer, 1993), and (Lewis, 1994).

Virzi (1992) used empirical data from three experiments and Monte Carlo simulation to conclude that problem discovery rate and the number of participants establish an asymptotic relationship. In his experiments, trained testers observed that the amount of discovered problems depends on the number of participants and the likelihood of discovering a problem. This last parameter is known as *problem discovery rate* and represents the average of the fraction of problems observed for each user (or the average of the proportion of users that detected each problem). The proportion of problems discovered was modelled with the cumulative binomial probability formula, as follows:

$$\text{Proportion of problems} = 1 - (1 - p)^n \qquad (1)$$

Where n is the number of participants, and p is the problem discovery rate.

The problem discovery rate, for a testing session is equal to the quotient between the number of unique problems detected, and the number of problem occurrences observed by all participants.

In (Nielsen & Landauer, 1993) data derived from eleven usability tests, and statistical modelling were used to reach a similar model. In this study, the number of detected problems is estimated as a function of n, p and the total number of problems N, as follows:

$$\text{Number of problems} = N\,[1 - (1 - p)^n] \quad (2)$$

Both approaches estimate the number of participants (n) required to uncover a goal percentage of problems, with a given problem discovery rate (p).

$$n = \log(1 - \text{Goal})/\log(1 - p) \quad (3)$$

Lewis (1994) applied the techniques used in (Virzi, 1992) to empirical data from usability testing conducted on a piece of software for office applications. The findings of this study coincided with Virzi's results. Nevertheless, he noted a potential overestimation of p in small-sample estimation.

Several authors have challenged the soundness of modelling problem rate discovery with a single value for p. Woolrych and Cockton (2001) contend that problems do not affect users uniformly, thus estimation based solely on problem frequency is misleading. Caulton (2001) argues that due to the heterogeneity of users, different types of users will discover different kinds of problems. Therefore, the model should incorporate a term that considers the number of user sub-groups. Turner, Lewis, and Nielsen (2006) responded to criticism of sample size formulae by providing a method to adjust the estimated average problem frequency.

2. Methodology

The steps involved in this study are summarized as follows:

- Obtain data from user testing.
- Process data to identify unique and repeated problems.
- Calculate parameters and metrics from this data to make observations.

One of the components of the last step in this process is using the calculated parameters to determine the number of user tests or samples required to achieve a level of problem discovery. Later, we employed a permutation test to investigate the distribution of the estimates of the problem discovery rate. In addition, we analyzed the accuracy of the estimation by comparing the mean scores of the estimates against the true value of p. Finally, this number was used to answer the question of whether the number of user tests undertaken was enough to reach a percentage of problem detection greater than 80%.

To review the consistency of the results, the parameters from two different datasets were computed.

2.1. Data sources

To obtain data required for the proposed analysis, two datasets were sourced from two independent rounds of testing of a web interface. In the tests, participants were requested to identify usability problems in the interface under study. In total, 34 different respondents participated in both surveys, 17 testers each. The second survey was undertaken two weeks after the first one.

To prepare the user testing data for parameter estimation, two passes were made over each dataset. The first pass was used to identify unique problems, which were catalogued and numbered. Later, the second pass counted which users identified which problem or problems. This process then resulted in a grid structure which shows the problem count for each user test, specifying the problem or problems observed by each tester. This process also allowed the identification of problems identified by more than one tester. Table 1 shows the processed data from Dataset #2

2.2. Parameter estimation

As per formula (1) the proportion of problems discovered depends on the problem discovery rate (p). To estimate this parameter, the quotient between the number of unique problems and the number of problem occurrences observed by participant is computed. As an example, consider the data in Table 1, in which there are 17 participants, 8 problems, and 28 problem observations (cells containing an "x"), with these values p is 0.21 = 28/(8*17).

2.3. Monte Carlo simulation

In many real-life applications, resource constrains prevent to gather enough participants to properly estimate the sample size. In such scenarios, p is estimated from small samples, using rules of thumb such as the "magic number 5". Hertzum and Jacobsen (2001), and Lewis (2000) investigated the effect of this practice, finding that small-sample estimation produce overestimation of this parameter, which will potentially lead to underestimate the required sample size, and to overestimate the proportion of problems discovered in a usability testing.

To illustrate this, suppose that in *Table 1* p is computed after the sixth test. The number of problems discovered up until that point is 3, with 7 problem occurrences. The value of p computed with these data would be 0.39 = 7/(3*6), which results in an overestimation of 85% of its true value. If the proportion of problems discovered were projected with the estimated of p, a practitioner would overestimate the number of problems uncovered and stop the testing earlier than needed to achieve a reasonable goal of problem discovery.

Since the selection of the sample to estimate p is arbitrary, different samples (data subsets) will produce different estimates. To investigate the distribution of the estimates, we used Python language and NumPy package to write a program that implements Monte Carlo sampling with 1000 permutations. According with (Lewis, 2000) this number of permutations produce a close

approximation to complete factorial combination. The outcomes of the permutation test were used to compute statistics of the distribution of p, and the proportion of problems discovered, as a function of the sample size across permutations.

Table 1. Problem count grid from data-set #2

Participant	Problem Id								Problem Count
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	
1	x	x							2
2									0
3									0
4		x							1
5	x	x							2
6		x	x						2
7	x					x			2
8		x			x				2
9	x	x							2
10	x				x			x	3
11	x			x	x		x	x	5
12	x	x					x		3
13		x							1
14	x								1
15									0
16	x								1
17		x							1
Problem count	9	9	1	1	3	1	2	2	28

Note. In the grid a cell containing an "x" denotes the observation of one of the problems by one of the users.

2.4. Adjustment of the estimation of p

In (Lewis, 2001) several adjustment techniques are reviewed and synthesized into one method, the author regarded as the most accurate, to adjust the estimate of p. The formula to adjust the value of the p, is:

$$p_{adj} = 1/2\ [(p_{est} - 1/n)(1 - 1/n)] + 1/2[\ p_{est}\ /(1 + GT_{adj})] \qquad (4)$$

Where n is the sample size used to compute the initial estimate of p, and GT_{adj} is the Good–Turing adjustment to probability space. GT_{adj} is obtained by dividing the number of problems that occurred once by the number of different problems. Going back to the example where p was estimated with a six-participant sample. The estimate for p applying Lewis' adjustment is 0.229, which compared to true p gives a deviation of 11% (significantly lower than that of the initial estimate).

3. Results

3.1. Parameter estimation

The number of users required to uncover a given percentage of usability problems can be projected as a function of the sample size used to estimate p. Tables Table ***2*** and Table ***3***, and figures Figure ***1*** and Figure ***2*** present the estimates of the sample size required to uncover 80%, 90% and 99% of problems computed from datasets 1 and 2.

Table 2. Projected sample size to achieve 80%, 90% and 99% of problem detection for Dataset #1

	Sample size															
Problem detection	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
80%	2	3	5	6	6	7	9	9	10	11	12	12	11	12	13	14
90%	2	4	6	8	9	10	12	13	14	16	17	17	16	17	19	20
99%	5	8	13	17	17	21	25	26	27	31	34	35	33	34	37	40

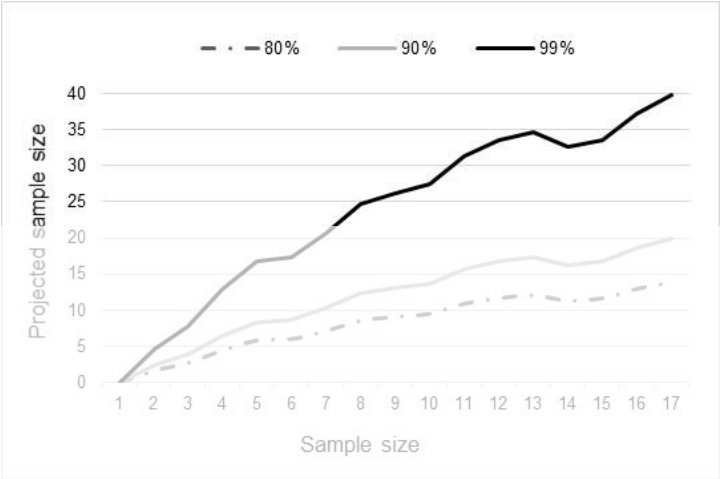

Figure 1. Required Sample Size estimated from Dataset #1.

Table 3. Projected sample size to achieve 80%, 90% and 99% of problem detection for Dataset #2

	Sample size															
Problem detection	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
80%	2	4	3	2	3	4	5	5	5	6	6	6	6	7	7	7
90%	3	6	5	3	5	6	7	7	7	8	8	8	9	9	10	10
99%	7	11	10	7	9	12	14	14	15	17	16	17	17	19	19	20

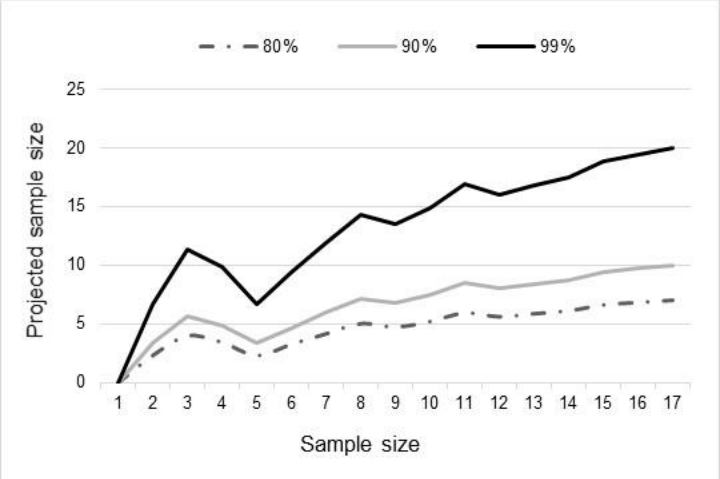

Figure 2. Required Sample Size estimated from Dataset #2.

3.2. Statistics of the distributions of the estimation of p

Monte Carlo simulation was undertaken to compute statistics of the distributions of the estimation of p, and the proportion of problems discovered for the two datasets as a function of sample size, across all permutations. The following statistics were computed:

a) *Means of the estimates of p*

Table 4 and *Figure 3* show the distribution of the means of the estimates of p as a function of sample size.

Table 4. Mean estimates of p for datasets 1 and 2

Sample size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Dataset #1	1.00	0.53	0.37	0.29	0.24	0.21	0.19	0.17	0.16	0.15	0.14	0.13	0.13	0.12	0.12	0.11	0.11
Dataset #2	1.00	0.62	0.48	0.42	0.37	0.34	0.31	0.29	0.28	0.26	0.25	0.24	0.23	0.22	0.22	0.21	0.21

Figure 3. Mean estimated p as a function of sample size

b) *Overestimation ratios of mean estimated* p

The overestimation ratios were computed against the true value of p of each dataset, and provide a measurement of the deviation of the means of the estimates as a function of sample size. A ratio of 1 indicates no overestimation, while a ratio greater than 1 denotes a overestimation percentage equal to (ratio – 1)*100.

Table 5. Overestimation ratios for datasets 1 and 2

Sample size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Dataset #1	9.15	4.83	3.39	2.67	2.23	1.95	1.74	1.58	1.47	1.37	1.29	1.23	1.17	1.12	1.08	1.04	1.00
Dataset #2	4.86	2.99	2.35	2.04	1.82	1.66	1.52	1.42	1.34	1.27	1.21	1.16	1.12	1.08	1.05	1.02	1.00

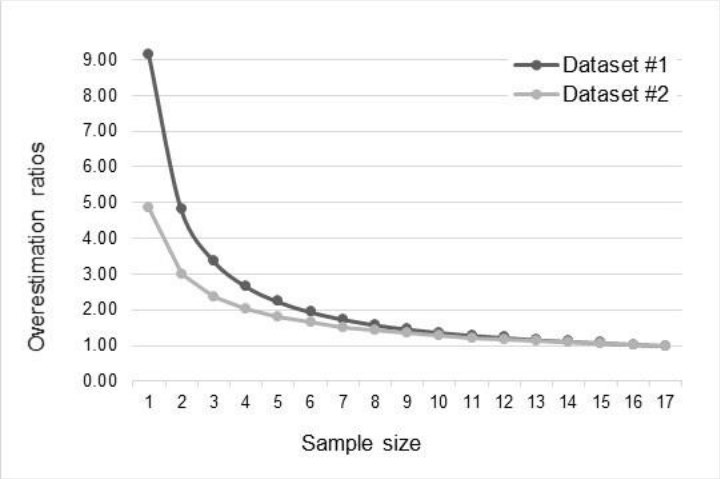

Figure 4. Overestimation ratios of p as a function of sample size

c) Root medium square error (RMSE) of mean estimated p

The root mean square error is the mean squared difference between each data point and the true value of p for the distribution. Compared to the standard deviation, the RMSE provides a more accurate measure of accuracy because it is sensitive to both the central tendency and variance. The lower the value the RMSE, the lower variance of the measurement. A RMSE of 0 indicates a perfect estimate.

Table 6. RMSE of mean estimated p for datasets 1 and 2

Sample size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Dataset #1	0.89	0.43	0.27	0.19	0.14	0.11	0.09	0.07	0.06	0.05	0.04	0.03	0.02	0.02	0.01	0.01	0.00
Dataset #2	0.79	0.43	0.30	0.24	0.19	0.15	0.12	0.10	0.08	0.07	0.05	0.04	0.03	0.03	0.02	0.01	0.00

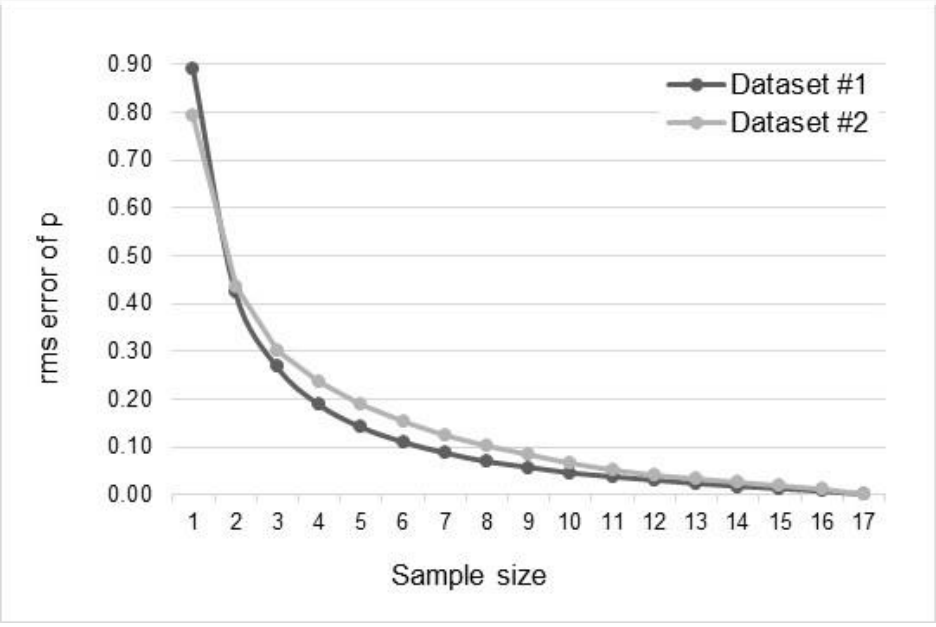

Figure 5. RMSE of estimated p as a function of sample size

3.3. Adjustment of the estimates of mean p

To adjust the estimates of the means of p, the method described in (Lewis, 2001) was used. Table 5 shows statistics of the distribution of adjusted p along with the same statistics of the distribution of the initial estimation.

Table 7. Statistics of the distribution of unadjusted and adjusted p for datasets 1, y 2

	DataSet #1			DataSet #2	
	Initial	Adjusted		Initial	Adjusted
median	0.27	0.16	median	0.27	0.07
average	0.34	0.18	average	0.16	0.09
mode	0.21	0.15	mode	0.11	0.06

3.4. Projected proportion of problems and sample for unadjusted, adjusted and true p

Figure Figure 6, *Figure Figure 7*, and *Table 8* show the proportion of problems discovered against the projected sample size for the average values of p (initial, adjusted, and true) for both datasets.

Figure 6. Proportion of problems discovered for Dataset #1 against projected p

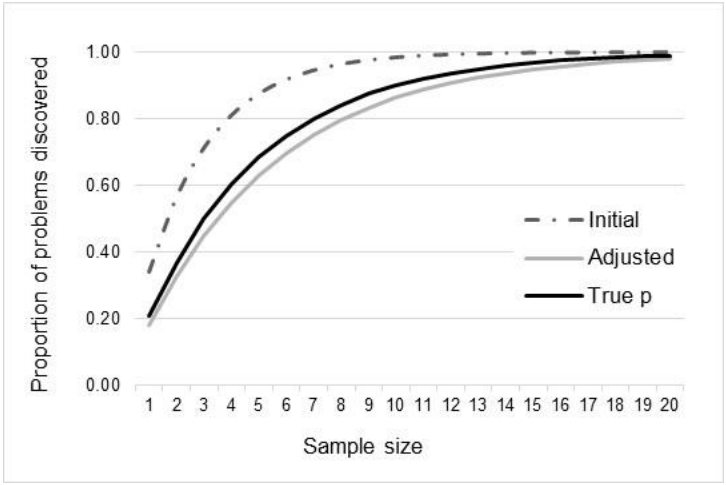

Figure 7. Proportion of problems discovered for Dataset #2 against projected p

Table 8. Proportion of problems discovered as function of p and sample size

	DataSet 2			DataSet 1		
Projected sample	Initial	Adjusted	True p	Initial	Adjusted	True p
1	0.34	0.18	0.21	0.16	0.09	0.11
2	0.56	0.33	0.37	0.29	0.17	0.21
3	0.71	0.45	0.50	0.41	0.25	0.30
4	**0.81**	0.55	0.60	0.50	0.31	0.37
5	0.87	0.63	0.68	0.58	0.38	0.44
6	**0.92**	0.70	0.75	0.65	0.43	0.50
7	0.95	0.75	**0.80**	0.70	0.48	0.56
8	0.96	**0.80**	0.84	0.75	0.53	0.61
9	0.98	0.83	0.87	0.79	0.57	0.65
10	0.98	0.86	**0.90**	**0.83**	0.61	0.69
11	**0.99**	0.89	0.92	0.85	0.65	0.72
12	0.99	**0.91**	0.94	0.88	0.68	0.75
13	1.00	0.92	0.95	**0.90**	0.71	0.78
14	1.00	0.94	0.96	0.91	0.73	**0.80**
15	1.00	0.95	0.97	0.93	0.76	0.83
16	1.00	0.96	0.98	0.94	0.78	0.85
17	1.00	0.97	0.98	0.95	**0.80**	0.86
18	1.00	0.97	0.98	0.96	0.82	0.88
19	1.00	0.98	**0.99**	0.96	0.83	0.89
20	1.00	0.98	0.99	0.97	0.85	**0.90**
21	1.00	0.98	0.99	0.97	0.86	0.91
22	1.00	**0.99**	0.99	0.98	0.87	0.92
23	1.00	0.99	1.00	0.98	0.89	0.93
24	1.00	0.99	1.00	0.98	**0.90**	0.94
25	1.00	0.99	1.00	**0.99**	0.91	0.95
26	1.00	0.99	1.00	0.99	0.91	0.95
27	1.00	1.00	1.00	0.99	0.92	0.96
28	1.00	1.00	1.00	0.99	0.93	0.96
29	1.00	1.00	1.00	0.99	0.94	0.97
30	1.00	1.00	1.00	0.99	0.94	0.97
31	1.00	1.00	1.00	1.00	0.95	0.97
32	1.00	1.00	1.00	1.00	0.95	0.98
33	1.00	1.00	1.00	1.00	0.96	0.98
34	1.00	1.00	1.00	1.00	0.96	0.98
35	1.00	1.00	1.00	1.00	0.96	0.98
36	1.00	1.00	1.00	1.00	0.97	0.98
37	1.00	1.00	1.00	1.00	0.97	**0.99**
38	1.00	1.00	1.00	1.00	0.97	0.99
39	1.00	1.00	1.00	1.00	0.97	0.99
40	1.00	1.00	1.00	1.00	0.98	0.99
41	1.00	1.00	1.00	1.00	0.98	0.99
42	1.00	1.00	1.00	1.00	0.98	0.99
43	1.00	1.00	1.00	1.00	0.98	0.99
44	1.00	1.00	1.00	1.00	0.98	0.99

Note. Highlighted rows indicate the projected sample size required to uncover 80%, 90%, 99% of problems

4. Discussion

4.1. Parameter estimation

There are two quite significant outcomes from the analysis of the parameter estimation (Figure 1 and Figure 2). The first brings into question the rules of thumb proposed by Virzi (1992), Nielsen and

Landauer (1993), and Lewis (1994), in the estimation of the number of user tests required to obtain the desired level of problem discovery.

The second significant outcome is the impact on the sample size of users required for increasing levels of problem discovery. As can be seen from the results from both data-sets, the amount of extra testing required to reach a 99% problem discovery outcome is significantly higher than that of the 90% level. This points towards a higher marginal cost of each extra problem discovered, especially once past the 90% threshold.

The analysis also confirms the commentary by Lewis (1994) on the variability problems associated with using small samples to estimate the number of samples needed for the required level of problem discovery. Even though the goal of Lewis' adjustments to the p value were to designed improve the calculation of the sample size required for smaller data-sets, the smaller user-test end of the scale for both data-sets showed quite a lot of variability between the test sets, however on both charts the results level out after about eight or nine user tests, thus providing a level of comfort in the suggested sample size results.

This eventual levelling out of results suggests a way to use this type of analysis in practice. If the percentage of problem discovery required is known, after each user test this analysis can be run to determine if more tests are required. For example, if 90% problem discovery is required, if this analysis had been done with Data-Set #1 then it would have been discovered that somewhere between 30 and 40 tests were required, thus pushing the testing past the actual 17 test users. Conversely, testing on Data-Set #2 could have stopped earlier as only 12 or 13 tests were required to hit the 90% problem discovery threshold.

In both scenarios, the question remains as to the "levelling out" point. However, this could be calculated using traditional statistical techniques.

4.2. Overestimation and adjustment of p

Analysis of the distribution of the estimates of p confirms the assertion in (Lewis, 2000) that the effect of overestimation is significant for small samples. Figure 3 shows that the means of the estimates increment their deviations exponentially as the size of the sample used to estimate decreases. In Figure 4, all samples less than 8 produce an overestimation greater than 50%, while in Figure 5 the RMS error grows consistently as the sample size shortens. Tables Table ***5*** and Table ***6*** confirm this trend. Considering the overestimation ratios, and the RMS errors of the mean estimates, it is evident that the smaller the sample the lower the accuracy of the estimation.

Comparison of the projected proportion of problem detection (and the associated sample size) computed with the initial estimate, the adjusted estimate, and true p (Table 6) shows that the initial estimate reaches a proportion greater than 80% earlier than the adjusted estimated, and the true value of the problem discovery rate. For dataset 1, the projected sample size to achieve 80% of

problems is 9 when computed with the initial estimate (5 testers less that the required sample calculated with true p), while the adjusted estimate projects a required sample of 17 users. On the other hand, for dataset 2, the difference between the projected sample computed with the initial estimates and the true sample size is lower (3 users). However, it is worth to note, that the error of estimation increases for 90% and 99% of problem detection. This can be observed in Figure 6 and Figure 7 where the adjusted estimation is closer to the true projection than the initial estimation is.

Several studies (Virzi, 1992), (Nielsen & Landauer, 1993), (Lewis, 1994), (Caulton, 2001), (Turner, Lewis, & Nielsen, 2006), (Hwang & Salvendy , 2010) suggest that when conducting user testing it is sufficient to use rules of thumb such as 4±1and 10±2 to estimate the number of users required.

In our study, we discovered that the use of such rules of thumb would have underestimated the actual number of users required to achieve reasonable levels of problem discovery. In the two data-sets studied, the number of users required to achieve 90% problem discovery were 20 and 10 respectively. This coincides with findings in (Bevan, et al., 2003), (Woolrych & Cockton, 2001), (Spool & Schroeder , 2001), (Faulkner, 2003), and (Lindgaard & Chattratichart , 2007). Furthermore, using small sample sizes can also be problematic as this produces large variability in testing results which cannot be fully adjusted for.

Since the potential costs of achieving problem discovery at the 99% level are significantly higher than that of achieving the 90% level, the use of these levels should be carefully considered unless the development is for applications for which the cost of problems is quite high.

One of the exciting possibilities of the results of this study would be the inclusion of continual problem discovery metrics into a user testing regime. Through continual testing of results, developers could optimize their testing to only include the required number of tests up to the desired problem discovery rate. This has the potential to concentrate testing resources on those cycles that need it rather than equally spreading resources across all cycles.

One possible scenario of use for the 4±1 rule is in time constrained agile cycles, more appropriately towards the start of projects. At this part of the project cycle, there is almost no point in discovering 100% or even 90% of possible problems if development is moving a pace which essentially wipes these problems out or replaces them with new ones.

If a project is severely budget limited, then rules of thumb such as 4±1and 10±2 will also come into play, although we would suggest that the 4±1 rule be only used on relatively simple projects.

This study was conducted in a specific project. However, the applied methodology could be applied to different project task types, different user groups and different environments to see if the rules of thumbs on the number of user tests required can be specified for different type of build projects of

differing complexity and task orientation. Such analyses should use at least eight or nine user tests in their parameter estimation to eliminate the small sample size issues that presented in this study.

Bibliography

Bevan, N., Barnum, C., Cockton, G., Nielsen, J., Spool, J., & Wixon , D. (2003). The "magic number 5": is it enough for web testing? In ACM (Ed.), *CHI '03 Extended Abstracts on Human Factors in Computing Systems (CHI EA '03)*, (pp. 698-699). Ft. Lauderdale, Florida, USA.

Caulton, D. (2001). Relaxing the homogeneity assumption in usability testing. *Behaviour & Information Technology, 20*(1), 1-7.

Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers, 35*(3), 379-383.

Hertzum, M., & Jacobsen, N. (2001). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction, 13*(4), 421-443.

Hwang, W., & Salvendy , G. (2010). Number of people required for usability evaluation: the 10±2 rule. *Commun. ACM, 53*(5), 130-133.

Lewis. (1994, June). Sample sizes for usability studies: Additional considerations. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 36*(2), 368-378.

Lewis. (2000). *Overestimation of p in problem discovery usability studies: How serious is the problem.* Tech. Rep.

Lewis. (2000). *Validation of Monte Carlo estimation of problem discovery.* Raleigh, NC: International Business Machines Corp.

Lewis. (2001). Evaluation of procedures for adjusting problem-discovery rates estimated from small samples. *International Journal of Human-Computer Interaction, 13*(4), 445-479.

Lindgaard, G., & Chattratichart , J. (2007). Usability testing: what have we overlooked? In ACM (Ed.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (pp. 1415-1424). San Jose, California, USA.

Nielsen, J., & Landauer, T. (1993). A Mathematical Model of the Finding of Usability Problems. In ACM (Ed.), *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, (pp. 206-213). Amsterdam, The Netherlands.

Spool, J., & Schroeder , W. (2001). Testing web sites: five users is nowhere near enough. In ACM (Ed.), *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, (pp. 285-286). Seattle, Washington.

Turner, C., Lewis, J., & Nielsen, J. (2006). Determining usability test sample size. *International encyclopedia of ergonomics and human factors, 3*(2), 3084-3088.

Virzi, R. (1992, August). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors: The Journal of the Human Factors and Ergonomics Society, 34*(4), 457-471.

Woolrych, A., & Cockton, G. (2001). Why and when five test users aren't enough. In C. Editions (Ed.), *Proceedings of IHM-HCI 2001 conference*, *2*, pp. 105-108. Toulouse, FR.